The use of safety suitability tests for the assessment of immigrant and majority group job applicants

te Nijenhuis, J.; van der Flier, H.

published in
International Journal of Selection and Assessment
2004

DOI (link to publisher)
10.1111/j.0965-075X.2004.277_1.x

document version
Publisher's PDF, also known as Version of record

The Use of Safety Suitability Tests for The Assessment of Immigrant and Majority Group Job Applicants

Jan te Nijenhuis*
Leiden Universiteit

Henk van der Flier
Vrije Universiteit

Safety jobs, meaning jobs where employees are responsible for the safety of customers, other employees and/or public in general are of special importance for our present society and deserve continued attention from I/O psychologists. The central question addressed in this study is whether the scores on safety suitability tests are comparable for immigrants and majority group members. Use was made of test data on first-generation immigrants ($N = 786$) and majority group members ($N = 584$) who applied for blue collar jobs at the Dutch Railways and at regional bus companies. The tests used measured selective attention, attentional speed, continuous attention, perceptual-motor ability and general mental ability. Immigrants' mean scores are systematically below the level of the mean scores of the majority group. The tests appear to have a strong dimensional comparability between the different groups. There is very little indication of test bias. The increasing number of immigrants and the increase of safety jobs pose challenges for selection psychologists. It is suggested that continued use of safety suitability tests is needed to keep the number of safety accidents at a minimum.

Introduction

Derailed trains and crashed planes tend to get front-page coverage in newspapers. Safety errors in public transportation are rare and rarely lead to accidents; however, accidents are counted in number of people died and injured. Safety jobs, meaning jobs where employees are responsible for the safety of customers, other employees and/or public in general, are of special importance for our present society and deserve continued attention from I/O psychologists. Proper selection of personnel is one of the keys to ensure that qualified people fulfill jobs in which safety plays an important role. Safety jobs are as varied as engine drivers, pilots, sewing machine operators, police, fire fighters, security guards, monitoring and surveillance jobs. The number of safety jobs is probably rising. Safety errors are costly, so testing for safety suitability may be a good investment. In the present study, we begin by describing testing for safety at Dutch Railways and then investigate whether tests used to measure the domain of safety suitability are biased against immigrant job applicants.

*Address for correspondence: Jan te Nijenhuis, Leiden University, Social and Organizational Psychology, PO Box 9555, 2300 RB Leiden, The Netherlands. E-mail: JanteNijenhuis@planet.nl

Safety Suitability

An important aspect of selection for specific positions with the Dutch Railways and urban and regional transport companies is safety suitability. Testing for safety suitability takes place when job applicants apply for jobs in which the safety of others or of the traffic is involved. With concern to psychological requirements, safety suitability is determined by the combination of cognitive functioning, attention functions, perceptual-motor abilities, and personality. So, safety suitability is not a unidimensional construct, but a description of a domain. Positions that require safety suitability include those of engine driver, guard, train traffic controller, bus driver, shunter, and railway station assistant. Although the specific tasks and behaviors involved in ensuring safety are different across these positions formal job analysis based on a Dutch version of the Position Analysis Questionnaire (McCormic, Jeanneret, & Mecham, 1972) revealed that the various jobs can be

characterized by a number of common safety dimensions or attributes indicating task requirements (Boomsma-Suerink, 1985). The most critical requirements were shown to be those in the perception and reaction fields, especially color discrimination, visual and auditory acuity, perceptual speed, spatial orientation, flexibility of visual form perception, selective and continuous attention, simple reaction time and speed in discriminate reaction. Color discrimination and visual and auditory acuity are checked during the medical examination. Perceptual speed, spatial orientation and flexibility of visual form perception are included in the General Aptitude Test Battery (see Method section of this paper). The so-called safety suitability tests were designed to measure selective and continuous attention, and speed in discriminate reaction (see also van der Flier, Schoonman, & Pouw, 1993).

The safety suitability tests are aimed at excluding risks to practical functioning and have moderate-to-reasonable predictive validity for various performance safety criteria as is shown in a substantial number of predictive validation studies, including a meta-analysis (Arthur, Barrett, & Alexander, 1991; Bukasa & Wenninger, 1985; Evers, van Vliet-Mulder, & ter Laak, 1992; Fletcher & Geary, 1993; Schoonman & Bosch, 1981; van der Flier & Schoonman, 1988; see Method for details) and have high face validity. Virtually all European railroad companies use safety suitability tests (International Union of Railways). The safety suitability tests of the present study or their computerized versions are or were also used in Great Britain, Austria, Germany, Sweden, and Hong Kong.

A high level of safety suitability is required for the position of engine driver because perceptual-motor and attention functions directly influence performance. Because of the need for strict adherence to time schedules, time pressure is high, and consequences of mistakes are grave and direct, whereas the possibility of support by others is limited. Moreover, sustained attentiveness is required, and both the work speed and the order of activities are largely not self-determined. Perception and perceptual-motor activities constitute important parts of the job. The essential task of the engine driver is to move a train from A to B; during this task the perception is focused on coded, well-described signals and psychomotor activities are restricted to a fixed number of operations: speed control by adjustment of the power of the engine and the brakes (Pouw, 1991). Learning the system of safety rules, including all the different signs, is quite difficult: it takes a long schooling program to master the safety skills. One of the hallmarks of good engine drivers is that they avoid making safety errors. Safety errors occur when engine drivers do not react (correctly) to the information from the environment; this information may consist of (1) coded signals from the cabin's dashboard or from signs alongside the railroad track, and (2) non-coded signals, such as failure of the computerized safety systems, a car on a level crossing, a cow on the track, or sudden, unexpected hail on the rails and the overhead wires, denoting unexpected or emergency situations. Luckily, the chances of safety error-related accidents are extremely small. van der Flier and Schoonman (1988) report about one case of passing through signals at danger (SPD case) for every million train-kilometers, and SPD cases resulting in accidents are rare. So, the great majority of safety errors do not lead to accidents.

Test Bias

Potential test bias against minority groups is a matter of concern for selection psychologists. In the US, blacks generally score on average a standard deviation lower on standardized intelligence tests than do whites. Much research was carried out to discover whether this was caused by test bias or whether it reflected a lower mean level of capacities in the black group (see Jensen, 1980). Schmidt, Ones, and Hunter (1992) state that the issue of fairness in testing has essentially been resolved as a research topic in personnel psychology in the US, and that because of this consensus, few new studies are being conducted. This conclusion was supported in 1994 by the opinion of 52 experts in intelligence and allied fields (Gottfredson, Arvey, Bouchard, Carroll, Cohen, Dawis, Detterman, Dunnette, Eysenck, Feldman, Fleishman, Gilmore, Gordon, Greene, Haier, Hardin, Hogan, Horn, Humphreys, Hunter, Itzkoff, Jackson, Jenkins, Jensen, Kaufman, Kaufman, Keith, Lambert, Loehlin, Lubinski, Lykken, Lynn, Meehl, Osborne, Perloff, Plomin, Reynolds, Rowe, Rushton, Sarich, Scarr, Schmidt, Schoenfeldt, Sharf, Spitz, Stanley, Thiessen, Thompson, Thorndike, Vernon, & Willerman, 1997) who state that one of the conclusions of mainstream science is that intelligence tests are not culturally biased against American blacks or other native-born, English-speaking peoples in the US but that IQ scores predict equally accurately for all such Americans, regardless of race and social class.

Research into testing of bilinguals (mostly Hispanics/Mexican-Americans) shows that below a certain level of English-language proficiency tests with a substantial language component become more of a measure of English-language proficiency than of *g* (Pennock-Román, 1992). However, the tests appear to still give an accurate short-term prediction of scholastic achievement and job proficiency (Jensen, 1980; Reynolds & Kaiser, 1990). Third World immigrants into the Netherlands are probably best compared to bilingual immigrants into the US Te Nijenhuis and van der Flier (1999) provide a comprehensive review of the extensive research on test bias against immigrant children and job applicants in the Netherlands from 1984 to 1999. Immigrants are mainly from Surinam, the Netherlands Antilles, Morocco, and Turkey. The empirical studies show that tests can be used within culturally homogeneous groups and can be used rather well for comparisons between immigrants and majority

group members; analyses of internal bias show that tests strongly measure the same dimensions and that the proportion of biased items is small; analyses of prediction bias show only little differential prediction. So, there is test bias, but its effects are not strong. Language bias is the only consistent and influential biasing factor found in cognitive tests and personality questionnaires when assessing immigrants (te Nijenhuis, Evers, & Mur, 2000; te Nijenhuis & van der Flier, 1997, 1999, 2000, 2001, 2003; te Nijenhuis, van der Flier, & van Leeuwen, 1997, 2003). When comparing test scores of people who do not have an adequate level of proficiency in the target language and bilinguals (i.e. most immigrants) with test scores of native speakers a distinction is usually made between verbal and non-verbal tests. Subtests with a substantial verbal component measure to an undesirable extent Dutch-language proficiency and underestimate the level of *g* of the tested non-native speakers. The more limited their proficiency in Dutch, the larger the underestimate. Using a mixture of culture-loaded and culture-reduced tests te Nijenhuis and van der Flier (2003) found that the highly verbal subtest Vocabulary of the GATB is so strongly biased that it depresses the score on Vocabulary with 0.92 *SD*, leading to an underestimate of *g* based on GATB IQ with as much as 1.8 IQ points due to this single-biased subtest alone, whereas the other 7 subtests combined show only very little bias and only underestimate *g* based on GATB IQ with 1.5 IQ points. So, the eight subtests of the GATB combined underestimate *g* based on GATB IQ with 1.8+1.5 = 3.3 IQ points, or .22 *SD*. However, one should not forget that subtests with a strong verbal component usually constitute only a small part of a test battery; due to the use of sum scores the strong bias in tests with a verbal component becomes diluted. Te Nijenhuis and van der Flier (1999) conclude that tests can be used quite well, though not perfectly, for the assessment of immigrants.

The majority of bias studies, however, have been carried out using cognitive tests. This makes sense as they are the best predictors of job proficiency (Salgado, Anderson, Moscoso, Bertua, & de Fruyt, 2003; Schmidt & Hunter, 1998). To the best of our knowledge, no bias research has been carried out in applied settings using safety suitability tests or comparable tests.

Research Question

The research question addressed in this study is whether the scores on safety suitability tests are comparable for immigrants and majority group members.

Method

Research Participants

This project made use of test data on first-generation immigrants and majority group members who applied for blue collar jobs at the Dutch Railways and at regional bus companies in the Netherlands from 1988 until 1992. The application process included a psychological examination, which took place at the Work Conditions Service Unit of the Dutch Railways in ten centers throughout the Netherlands. The immigrant sample comprised the complete population of first-generation job applicants. A representative, random sample was selected from the complete population of majority group job applicants in such a way that the distribution with respect to the jobs and regions in this subsample was as close as possible to that in the immigrant group; Table 1 shows the distributions of the groups in terms of demographic variables. The data from the sample that took one of the safety suitability tests, the Groeps-bourdon, are reported, because this is the largest group.

Table 1. Distribution of immigrant group members and majority group members with respect to native country, size of group or subgroup, percentage of males, and age

Native country	*n*	% males	Mean age
Surinam	370	77.9	29.3
Netherlands Antilles	96	81.3	29.9
North Africa	103	96.8	26.2
Turkey	217	96.0	23.3
Netherlands	584	89.0	28.8

Note: The group of North Africans consists of persons mainly from Morocco, and also some from Algeria, Tunisia, Libya, and Egypt.

Tests

To describe the tests, use was made of Carroll's (1993) hierarchical intelligence model; *g* or general mental ability is at stratum III, broad abilities are at stratum II, and the narrow abilities are at stratum I.

ADM. The ADM (Attention Diagnostic Method; Rutten & Block, 1976) measures selective attention at stratum I and Attentional Speed at stratum II. The ADM consist of a series of numbers that are placed in random order on a screen. In part one, the candidate has to look up the numbers in the order from 11 to 59. When the number is found, the candidate has to give its value and color. In part two, the same numbers are given in a different random order and with different colors. Now there are smaller sized numbers under the main numbers. Again, the candidate has to look up the numbers in the order from 11 to 59. When the number is found, the candidate has to give the value of the smaller sized number and the color of the main number. The ADM consists of a fluorescent board with numbers on both sides; different sides of the board being used for part one and part two. The test is administered in a darkened room. The test has moderate predictive validity (Schoonman & Bosch, 1981). The number of mistakes gives the

best prediction for accident-related criteria. Arthur *et al*. (1991) report in their meta-analysis that selective attention tests give a moderate prediction of vehicular accident involvement.

Groepsbourdon. The *Groepsbourdon* [Group Bourdon] (GRBD, adaptation of the Bourdon-Wiersma test by Boomsma & Bosch, 1978) measures both selective attention and continuous attention, but most strongly selective attention at stratum I and Attentional Speed at stratum II. The GRBD consists of five pages with each ten lines of 25 figures of 3, 4, or 5 dots. The candidate has to cross out all figures with four dots as quickly as possible. Time used, No. mistakes, and No. omissions are registered. Traditionally, No. completed is taken as a measure of speed and No. omissions as a measure of accuracy. One can also look at variability in speed and accuracy, although no measures are available for them. A positive score consists of a large number completed with a small number of omissions. Candidates who combine small No. completed with large No. omissions are regarded as unfit.

A review of test research in the Netherlands by Evers *et al*. (1992) showed that the Bourdon–Vos test (Vos, 1988), which is related to the Groepsbourdon, has acceptable validity. The meta-analytic findings of moderate predictive validity for selective attention tests (Arthur *et al*., 1991) also apply to the Groepsbourdon.

DTG. The DTG (Determinations Gerät) [Determination apparatus] is a perceptual-motor test. It measures choice reaction at stratum I and perceptual-motor ability at stratum II. The traditional description of the test is that it measures ability to react, sensorimotor coordination ability, and precision of reactions; the capacity limit with regard to sustained, accurate reactions to variable visual and acoustic stimuli can also be determined, but this was not done in the present group. In non-systematic order, visual and acoustic stimuli are presented to which specific reactions must be given. The visual stimuli are presented on a screen, the acoustic stimuli are presented over a headphone. The reactions consist of pressing buttons on the reaction screen with the fingers and using pedals with the feet. The visual stimuli consist of five differently colored lights that appear on different places on the screen. A correct reaction consists of pressing the button with the same color. Two other visual stimuli are fixed yellow lamps on the left and right sides of the screen. A correct reaction consists of pressing the left or the right pedal. The acoustic stimuli consist of low and high tones. A correct reaction consists of using the left or the right black buttons on the reaction screen.

The test consists of three time-driven parts, in which the intervals between the stimuli are 1.1, 1.0, and 0.8 s (DTG1.1TD, DTG1.0TD and DTG0.8TD, respectively), and two reaction-driven parts (DTG1RD and DTG2RD, respectively) of each 150 s. Four performance measures are registered: No. correct, No. late, No. mistakes, and No. omissions (van Drie & Schoonman, 1993).

The scores on the DTG correlate moderately with accident criteria; the number of mistakes gives the best prediction of safety errors (van der Flier & Schoonman, 1988). This kind of study probably gives an underestimate of predictive validity, because criteria like "accidents caused" and "passing stop signals at danger" have a low frequency and are sometimes registered unreliably. Other predicitive validity data are reported by Fletcher and Geary (1993) and Bukasa and Wenninger (1985). When interpreting the scores on the DTG, experienced selection psychologists at the Dutch Railways look especially at the No. correct and the No. mistakes. The No. omissions and the No. late play a less important role for the interpretation of the test scores.

GATB. The GATB 1002 B (General Aptitude Test Battery) is a test of general intelligence. At stratum I Three-Dimensional Space measures Visualization, Vocabulary measures Induction and Lexical Knowledge, Arithmetic Reason measures Quantitative Reasoning, Computation measures Numerical Ability, Tool Matching measures Perceptual Speed, Form Matching measures Spatial Relations, Name Comparison measures Perceptual Speed and Numerical Ability, and Mark Making measures Aiming. At stratum II, the subtests Three- Dimensional Space, Tool Matching, and Form Matching measure Broad Visual Perception. The subtests Vocabulary and Arithmetic Reason measure both Fluid and Crystallized Abilities. The subtest Computation measures Crystallized Abilities. The subtest Name Comparison measures both Broad Visual Perception and Crystallized Abilities. The subtest Mark Making measures General Psychomotor Speed. The influential review of test research in the Netherlands by Evers *et al*. (1992) showed that the Dutch version of the test has good predictive validity, content validity, and construct validity.

Statistical Analyses

Descriptive Statistics. Mean scores were computed for immigrants and majority group members. The *g* score of research participants was computed by summing the products of participant's *z* scores and the subtest's *g* values for all the subtests. The *g* loadings were computed, using the first unrotated factor of a principal-axis factor analysis (Jensen & Weng, 1994). Because of the limited sampling of broad abilities of the GATB, it is not optimal for a precise and theoretically sound estimate of *g* loadings. The best estimate of the *g* loadings was found in a factor analytic study of the Dutch version of the GATB 1002 A with a large number of other tests, using the first unrotated factor of a principal axis factor analysis (Dutch GATB Manual; van der Flier & Boomsma-Suerink, 1994, p. 51).

The deviation of the mean scores of the immigrants on the subtests from the mean scores of the majority group members was expressed in terms of the standard deviation of the majority group.

Dimensional Comparability. The dimensional comparability of the tests for the majority group and the immigrant groups was examined by means of structural equation modeling, using EQS (Bentler, 1989). A model was used including the most relevant variables. When analyzing the correlations between the variables No. correct, No. late, No. mistakes, and No. omissions of the time-driven part of the DTG, it is problematic that the variables are not experimentally independent, meaning that they are not based on independent observations. Nunnally (1978, p. 434) and Gorsuch (1983, p. 301) warn against the use of experimentally dependent measures on the grounds that these may lead to spurious correlations. Because for every subtest the variables always add up to 100, the influence of third variables, such as speed/accuracy trade-off, learning effects, or tiredness may influence the correlations. If the No. correct becomes larger, the scores on the other three variables become, on average, lower so that an artificial negative correlation between No. correct and the other variables arises. Common error components also lead to spurious correlations. The variables, as regard to the different parts, are based on independent observations, so that comparing correlations by structural equations modeling (and also by factor analyses) is justified.

Several models with increasing degrees of constraint were fitted to the data. The following tests were examined: (a) tests of comparability of covariance matrices, (b) tests of the same number of factors in two groups, and (c) tests of the equality of factor loadings in two groups. The factor model tested across groups was te Nijenhuis and van der Flier's (2002) hierarchical model including cognitive and safety suitability tests, confirmed on the group of majority group applicants of this study. The model has a hierarchical factor structure. At the top of the hierarchy is *g* with two lower-order factors from the paper-and-pencil cognitive tests and two lower-order factors from the safety-suitability tests. All tests are influenced by the *g* factor. The first, cognitive factor is a hybrid of Fluid and Crystallized Intelligence and is called g_h. The second, cognitive factor is related to Broad Visual Perception and is called g_v. The third factor is called Attentional Speed and the fourth factor is called Perceptual-Motor Ability.

When working with large samples, even small differences between groups can lead to large chi-square values; these χ^2 values will make the small differences significant. For that reason, various researchers have suggested additional goodness-of-fit measures, such as the comparative fit index (CFI) (Bentler, 1989), which has been shown to be less susceptible to the effects of sample size than other measures.

Group Differences in Learning Effects and Speed/Accuracy Trade-offs. On the DTG, it was checked whether learning effects were present, and if so, whether they were stronger for immigrants. Score profiles of immigrant groups were compared with the score profiles of the majority group. The Groepsbourdon was checked for speed/accuracy tradeoffs.

Test Scores Tested for Jensen Effects. Spearman's hypothesis (Spearman, 1927) holds that black/white group differences in mean scores on cognitive tests are dependent on the *g* loading of the tests and not only on cultural variables or other hypothetical variables. This implies that the differences in scores between groups are larger as the *g* loadings of tests are higher. Rushton (1998) proposed that when a positive correlation occurs between *g* loadedness and variable *X*, the result be termed a "Jensen effect", because otherwise there is no name for it, only a long explanation of how the effect was achieved. So, the use of the term "Spearman's hypothesis" may be restricted to research in the US with Black and White groups, and can be seen as a special case of the general Jensen effect. When Jensen effects are found the conclusion is that differences in safety suitability test scores and GATB scores are attributable to group differences in mean intelligence level. Jensen (1993) specifies seven methodological requirements for the testing of Spearman's hypothesis and Jensen effects; they were all followed. To test the hypothesis that the correlation between the *g* loadings and the effect sizes is not caused by differences in reliabilities between the subtests, the *g* loadings and the standardized score differences were corrected for unreliability. Jensen (1985) advises using test-retest reliabilities. Bosch and Schoonman (1982) report test–retest reliabilities for the DTG and the ADM.

Results

To reduce the number of comparisons in the main analyses and to diminish the risk of accidental deviations or deviations of little practical significance, it was checked whether the data from the Turks and the North Africans, on the one hand, and the data from the Surinamese and the Antilleans, on the other, could be combined. These combinations seemed obvious, considering the similarity between the Surinamese and Antilleans with respect to their proficiency in Dutch. Through education and exposure to the media in their native countries, the Surinamese and Antilleans came into contact with the Dutch language and culture, which is not true of the Turks and North Africans. It was checked whether the correlation matrices of two groups were equal and it was found that there were no systematic differences for the compared groups.

Descriptive Statistics

ADM. Table 2 shows that immigrant groups have less favorable mean scores on both parts of the ADM than the majority group. If the mean scores of immigrant groups were conceived as the scores of a person, experienced selection psychologists at the Dutch Railways would have strong reservations with regard to the safety suitability,

Table 2. Mean time used in seconds and SDs by group on the ADM

	Group		
Subtests	Majority	Surinamese+ Antilleans	North Africans+Turks
ADM1			
Time used	311	352	397
SD	90	104	116
ADM2			
Time used	428	496	524
SD	121	128	131

Note: Majority group: $N = 371$; Surinamese: $n = 176$; Antilleans: $n = 35$; North Africans: $n = 52$; Turks: $n = 66$.

based on their score profiles. In combination with a moderate mean scores on the DTG, such an applicant would be rejected as presenting too large a safety risk.

Groepsbourdon. Table 3 shows that the mean difference in No. completed between the majority group and immigrant groups are a few percentage points and no more than a third of a standard deviation. Experienced selection psychologists are of the opinion that the differences in mean No. completed between the groups have a negligible effect on the prospects of being hired. A difference in No. completed of a few percentage points means that the distribution of the various groups over the deciles of the norm group hardly differs.

Table 4 shows that the mean total No. omissions in the immigrant group is 31–51%, or 0.34–0.56 *SD*s higher than in the majority group. Experienced selection psychologists see these differences as exerting an important influence on prospects of being hired. The differences found here imply that the immigrants generally score one or two deciles lower.

DTG. Table 5 shows that immigrant groups in all cases have a negative mean score profile on the DTG. The mean No. correct is always lower than in the majority group, whereas mean No. late, mean No. mistakes, and mean No. omissions is in most cases larger than in the majority group.

In immigrant groups both the mean No. correct and the mean No. mistakes are less good than the mean of the norm group. If the mean scores of the immigrant groups were conceived as the score profiles of a person, experienced selection psychologists at the Dutch Railways would assess these score profiles in the range just sufficient to moderate. The excessive No. mistakes plays an important part in this case. If the results on the personality tests and the intelligence tests were acceptable, these job applicants would in principle be hired, but with strong reservations with regard to safety suitability. Additional moderate assessments on the other safety suitability tests increase this reservation, which generally implies that during the probationary period a critical assessment of the practical functioning would take place.

GATB. The mean *g* score of the majority group was 2.06 ($SD = 2.66$); for the group of Surinamese and Antilleans it was -0.87 ($SD = 2.86$; $D = 1.10$), and for the Turks and North Africans it was -2.20 ($SD = 2.85$; $D = 1.60$). So, the mean differences in cognitive abilities between the groups were large.

Comparability of Dimensions

The comparability of the dimensions of the tests for the majority group and the immigrant groups was investigated by means of structural equation modeling (EQS). The covariance matrix of the majority group was compared, in separate analyses, with the covariance matrices of the Surinamese and the Antilleans, and the North Africans and

Table 3. Means and SDs of No. completed on the five parts of the GRBD by group

	Group					
	Majority		Surinamese+ Antilleans		North Africans+Turks	
Variable	*M*	SD	*M*	SD	*M*	SD
No. completed						
Part 1	219	41	217	46	225	44
Part 2	215	38	219	42	224	40
Part 3	213	37	222	43	225	42
Part 4	216	36	224	41	227	40
Part 5	221	38	230	42	233	40
Total	1083	180	1113	197	1133	189

Note: Majority group: $N = 584$; Surinamese: $n = 370$; Antilleans: $n = 96$; North Africans: $n = 103$; Turks: $n = 217$.

Table 4. Mean No. omissions and SDs by group and by part of the GRBD; also total No. omissions

	Group					
	Majority		Surinamese+ Antilleans		North Africans+Turks	
Variable	*M*	SD	*M*	SD	*M*	SD
No. omissions						
Part 1	2.4	2.9	3.0	3.5	3.5	4.1
Part 2	2.6	2.9	3.7	4.5	3.9	4.4
Part 3	2.9	3.1	4.3	4.9	4.4	4.4
Part 4	3.0	3.6	4.1	4.2	4.6	4.6
Part 5	3.0	3.2	4.7	4.8	4.7	4.5
Total	13.6	12.5	17.8	14.2	20.6	16.9

Note: Majority group: $N = 584$; Surinamese: $n = 370$; Antilleans: $n = 96$; North Africans: $n = 103$; Turks: $n = 217$.

Table 5. Mean scores and SDs by group on the DTG

	Group					
	Majority		Surinamese+ Antilleans		North Africans+Turks	
Subtest	*M*	SD	*M*	SD	*M*	SD
DTG1.1TD						
No. correct	79.3	15.6	69.2	18.5	72.9	15.6
No. late	11.7	9.1	15.0	10.2	12.5	8.0
No. mistakes	8.4	7.3	13.7	10.8	13.1	9.1
No. omissions	2.3	4.5	3.8	6.5	3.3	5.2
DTG1.0TD						
No. correct	70.7	19.8	60.5	21.1	65.1	18.0
No. late	17.8	12.4	21.5	13.2	17.3	11.1
No. mistakes	8.9	8.1	13.8	11.0	14.3	9.0
No. omissions	4.1	6.5	5.8	8.7	4.8	7.0
DTG0.8TD						
No. correct	29.4	19.4	22.2	15.7	28.5	17.3
No. late	43.5	13.2	44.6	12.4	40.0	12.1
No. mistakes	12.3	8.3	17.2	11.3	16.7	10.3
No. omissions	15.3	13.2	16.3	13.8	14.9	13.3
DTG1RD						
No. correct	165.9	22.0	151.6	33.6	157.0	28.5
No. mistakes	9.7	10.9	16.2	18.8	16.7	17.3
DTG2RD						
No. correct	165.1	21.2	154.0	26.7	160.8	20.4
No. mistakes	9.7	10.5	15.1	18.9	14.4	15.8

Note: Majority group: $N = 464$; Surinamese: $n = 221$; Antilleans: $n = 38$; North Africans: $n = 84$; Turks: $n = 160$.

the Turks. The values of the CFI were, respectively, .916 and .940. From this it may be concluded that the covariance matrices in the different comparisons are highly comparable.

The fit was further explored by fitting increasingly constrained models to the data. In the first analysis, a test of the same number of factors was examined. Because a hierarchical model was tested, the factors were oblique.

The data showed a good fit for the comparison of the majority group and the Surinamese/Antillean group, χ^2 (192, $N = 699$) = 599.34, $p < .001$, CFI = .920; and for the comparison of the majority group and the North African/Turkish group, χ^2 (192, $N = 638$) = 670.86, $p < .001$, CFI = .906. In the second analysis, the factor loadings were held equal across the groups. The fit deteriorated somewhat, but was still acceptable: for the comparison of the majority group and the Surinamese/Antillean group, χ^2 (210, $N = 699$) = 693.30, $p < .001$, CFI = .905; and for the comparison of the majority group and the North African/Turkish group, χ^2 (210, $N = 638$) = 733.30, $p < .001$, CFI = .897. In sum, the fit of the model postulating the same numbers of factors for the different groups was adequate and a model of equal factor loadings was also adequate, albeit slightly less so. The dimensions of the subtests for the majority group and the immigrant groups were highly comparable.

Differences in Learning Effects and Speed/Accuracy Tradeoffs

Groepsbourdon. In the majority group the No. completed remains as good as constant, whereas the No. omissions clearly deteriorates. In the immigrant group the No. completed improves slightly; expressed in standard deviations, however, this is a marginal improvement. The No. omissions shows a clear deterioration; in comparison with the deterioration in the majority group, the score of the immigrants deteriorates about 2 times as badly. Because of the functional relationship between No. completed and No. omissions it is, however, improbable that there is a learning effect for No. completed but not for No. omissions. It therefore appears more probable that the marginally improved No. completed is largely, and the highly deteriorated No. omissions is for a small part, caused by a small change in speed/accuracy trade-off.

The first alternative explanation for the large No. omissions in the immigrant group is that a linear increase in No. completed comes with an exponential increase in No. omissions. Regression analyses show that adding an exponential term in both groups does not improve the predictions with the linear model (for the majority group R^2 change = .00, $F = .78$, $p = .38$; for the immigrant group R^2 change = .00, $F = 2.9$, $p = .09$), so that the first alternative explanation is not supported.

The second alternative explanation is that immigrants put more emphasis on speed at the expense of accuracy. The comparison of scores between individuals and groups on the safety suitability tests is complicated by the fact that No. completed and No. omissions are not independent. A larger No. completed can be achieved at the expense of a larger No. omissions and a small No. omissions can be achieved by working slowly. To improve comparisons between scores of individuals, the scores on No. completed and No. omissions should be corrected for an extra emphasis on speed or an extra emphasis on accuracy.

Pew (1969) describes a methodology in which candidates are asked to vary their emphasis on speed and on accuracy during the tests, so that a systematic variation in mean reaction times and mistakes appears. Pew shows that reaction time relates linearly with the logarithm of No. correct/No. mistakes for an individual, but does not give unequivocal information about how this might lead to an improved comparison between scores of individuals. Later researchers like Meyer, Irwin, Osman, and Kounios (1988) have likewise failed to find an acceptable solution to this problem.

When looking at the relations between No. completed and No. omissions, it appears that three independent processes are at work: (1) capacity differences in concentration ability, (2) differences in speed/accuracy trade-off, and (3) a functional relation between No. correct and No. omissions.

The capacity differences cause good candidates to have a large No. completed and a small No. omissions and poor candidates to have a small No. completed and a large No. omissions. An emphasis on speed goes at the expense of more omissions and a stress on accuracy goes at the expense of the No. completed. The functional relationship between No. correct and No. omissions leads to the logical consequence that a higher No. completed comes with a higher No. omissions. If the likelihood of omissions is seen as being constant, the probability of omissions increases with the No. completed. The two measures are not independent.

Process 1 results in a negative correlation and processes 2 and 3 result in a positive correlation between No. correct and No. omissions within a group. The resultant of these three different processes, the correlation within a group, is an empirical fact. No research is known about the influence of the various processes on the size of the correlation. A correlation at the group level does not therefore appear to be a useful measure with which to answer the question of individual differences in speed/accuracy trade-offs. For this reason no conclusions can be drawn with regard to group differences in speed/accuracy trade-offs.

DTG. Table 5 shows that, broadly, the score profile of the immigrants improves in comparison with the majority group as the test continues. During the time-driven parts the No. correct, No. late, and No. omissions come closer to the mean of the majority group. The mean No. mistakes as regard to the three time-driven parts does not show a clear downward trend for every group in comparison with the majority group. However, there is a slight score improvement for the immigrant group as a whole. This profile improvement in comparison with the majority group may be explained by a change either in speed/accuracy trade-off or by a learning process. When there is greater emphasis on speed, a higher No. correct would be accompanied by a higher No. mistakes and No. omissions. When there is a

learning process, a higher No. correct would be accompanied by a lower No. mistakes and No. omissions. The data give indications that this is a learning process. For the reaction-driven part, the mean scores also come closer to the mean of the majority group. In comparison with the mean score on the first reaction-driven part, the mean No. correct on the second part is slightly improved and the No. mistakes is slightly decreased. This score improvement may be interpreted as a learning effect.

The interpretation of the data is complicated by two points that appear inconsistent. On the one hand, the factor solutions of the groups are highly comparable, which would point to the interpretation that differences in speed of learning play a comparable role in the groups. On the other hand, the score improvement in comparison with the majority group takes place in practically all cases, which would lead to an interpretation in terms of a learning effect. A possible explanation of these inconsistent findings is that the score improvement of immigrants partly consists of a constant. This constant may be interpreted as a "catch-up effect" that is equally large for all immigrants. Beside this general catch-up effect, individual differences in learning speed would appear to the same degree in both groups. A constant does not influence correlations, so it could be an explanation for the comparable factor structures in both groups. This interpretation in terms of a catch-up effect is supported by the fact that the immigrant job applicants have lower average scores on *g*. When tests are repeated, low-IQ participants improve their scores more rapidly than high-IQ research participants (Ackerman, 1987; Adams, 1957; Reynolds & Adams, 1954); when gain scores are taken as measures of learning ability, this leads to the erroneous conclusion that the persons with the highest intelligence have the worst learning ability. This pattern is consistent with the power law of practice (Anderson, 2000), which describes the common learning-curve phenomenon that, as individuals approach asymptotic performance, performance gains are increasingly difficult to obtain. Higher-ability individuals are approaching asymptotic performance earlier in task practice than lower-ability subjects. So, care should be taken with the interpretation of gain scores as learning effects, as the groups strongly differ in mean intelligence level. So, it cannot be concluded that there is a stronger learning effect for immigrants.

Safety Suitability Tests Tested for Jensen Effects

Experimentally independent variables were used to test for Jensen effects. A choice was made for ADM1TIME, ADM2TIME, DTG1.1TD No. correct, DTG1.0TD No. mistakes, DTG0.8TG No. omissions, DTG1RG No. correct, DTG2RG No. mistakes, because this set of variables seems a good representation of the ADM and the DTG. Because group differences in speed/accuracy trade-offs in the GRBD cannot be ruled out, a conservative strategy of not including these data was followed.

Tables 6 and 7 show that the variation in *g* loadings might be called good, that is, .14–.56 for the majority group and .19–.59 for the immigrants. The *g* loadings are highly comparable for both groups, so they may be averaged for the testing of the two groups combined. The Pearson correlation between the mean *g* and the standardized mean score difference is .22 for the Surinamese and the Antilleans, and .70 for the North Africans and the Turks. So, the difference in means on the safety suitability tests between the majority group and the immigrant group can be predicted with a high degree of precision from the *g* loadings, using the regression formula: effect size = $1.34g+.04$ for the Turks and North Africans; for the Surinamese and the Antillans effect size = $.38g+0.39$.

The standardized score differences were corrected for unreliability (see Table 8) to test the alternative hypothesis that the correlation between the *g* loadings and effect sizes is caused by differences in reliability; the standardized mean score differences were divided by the root of the reliability for each variable. As several test–retest reliabilities were below .60, correcting for unreliability would lead to overcorrections. We therefore choose a minimum value of .60 when correcting the effect sizes for unreliability. The Pearson correlation between the mean *g* loadings and the corrected standardized mean score differences is .14 for the Surinamese and the Antilleans, and .65 for the North

Table 6. Correlations of variables of the ADM and the DTG with the *g* score of the GATB by group, mean *g* loadings, and the standardized score difference in SDs between majority group and Surinamese+Antillean group (effect size)

Variable	*r* majority	*r* immigrants	Mean *g*	Effect size
ADM1TIME	−.54	−.47	.51	.46
ADM2TIME	−.56	−.51	.54	.56
DTG1.1TD No. correct	.40	.37	.39	.64
DTG1.0TD No. mistakes	−.34	−.20	.27	.61
DTG0.8TD No. omissions	−.25	−.32	.29	.08
DTG1RD No. correct	.39	.43	.41	.83
DTG2RD No. mistakes	−.14	−.21	.18	.52

Note: A positive effect size implies a less good mean score for immigrants.

Table 7. Correlations of variables of the ADM and the DTG with the *g* score of the GATB by group, mean *g* loadings, and the standardized score difference in SDs between majority group and North African+Turkish Group (effect size)

Variable	*r* majority	*r* immigrants	Mean *g*	Effect size
ADM1TIME	−.54	−.59	.57	.96
ADM2TIME	−.56	−.60	.58	.79
DTG1.1TD No. correct	.40	.45	.43	.41
DTG1.0TD No. mistakes	−.34	−.19	.27	.67
DTG0.8TD No. omissions	−.25	−.23	.24	.03
DTG1RD No. correct	.39	.44	.42	.59
DTG2RD No. mistakes	−.14	−.24	.19	.45

Note: A positive effect size implies a less good mean score for immigrants.

Table 8. Reliability coefficients of the variables

Variable	
ADM1TIME	.70
ADM2TIME	.73
DTG1.1TD No. correct	.69
DTG1.0TD No. mistakes	.43
DTG0.8TD No. omissions	.57
DTG1RD no. correct	.57
DTG2RD No. mistakes	.36

Africans and Turks. The results of these analyses demonstrate that the finding of Jensen effects is not an artifact of variation in reliability of the different parts of the safety suitability tests.

GATB Scores Tested for Jensen Effects

GATB data were tested for Jensen effects. Analyses were carried out in the approved manner (see te Nijenhuis & van der Flier [1997] for details). The Pearson correlation between the mean *g* and the standardized mean score difference is .67 for the Surinamese and the Antilleans, and .77 for the North Africans and the Turks. So, the difference in means on the subtests between the majority group and the immigrant group can be predicted with a high degree of precision from the *g* loadings, using the regression formula : effect size = $1.33g - .02$ for the Surinamese and the Antillans, and effect size = $2.22g - .19$ for the Turks and North Africans. The standardized score differences were corrected for unreliability and again we choose a minimum value of .60 when correcting the effect sizes for unreliability. The Pearson correlation between the mean *g* loadings and the corrected standardized mean score differences is .65 for the Surinamese and the Antilleans, and .76 for the North Africans and Turks. Again, finding Jensen effects is not an artifact of variation in reliability of the different parts of the various subtests.

Discussion

This study focused on the comparability of scores for immigrants and majority group members on safety suitability tests. For the immigrants, the mean scores on the ADM, DTG, and Groepsbourdon are systematically below the level of the mean scores of the majority group. The tests appear to have a strong dimensional comparability between the different groups. There is very little indication of test bias.

Test Bias

On the GRBD there are no large group differences in the mean No. completed, but for the mean No. omissions these differences do exist. It is not clear from the analyses whether the No. omissions in the immigrant group has a different meaning. The score profiles of immigrants on the DTG are negative: a low mean No. correct goes with a high mean No. mistakes. The mean scores improve, in comparison with the mean scores of the majority group, as the test progresses. These improvements are probably not attributable to differences in strategies and there is also lack of convincing proof that they are caused by a stronger learning effect in the immigrant group. It is possible that immigrants' score improvement is not complete after the five parts of the DTG; the available data do not allow conclusions. There are substantial to large mean group differences on the ADM.

Immigrants have, on all safety suitability tests, score profiles that are poorer than those of the majority group. There are no indications of group differences in meaning for the DTG and the ADM, but the No. omissions on the GRBD may have a slightly different meaning. The analyses give no indications of substantial differences in construct validity of the three tests. The question of whether the lower level of accuracy during the test translates into a lower level of concentration ability in an actual work situation can only be answered by a predictive validity study, but it is most plausible to assume that this will be the case.

Group differences on the safety suitability tests are related to their *g* loading. So, the more strongly the safety tests are correlated with *g*, the lower the mean scores of the immigrants are.

In sum, the scores of majority group members and immigrants appear to be highly comparable on safety suitability tests.

Limitations of This Study

An empirical study on differential prediction is needed to convincingly answer the question whether low safety suitability test scores for immigrants translate into low job performance. However, numerous empirical studies on test bias and differential prediction – using various tests – show no or only very limited bias, with the exception of studies on people who do not have a desirable level of proficiency in the target language and bilinguals when using tests with a substantial verbal component (see Schmidt *et al*., 1992; te Nijenhuis & van der Flier, 1999). So, it could be argued that the default hypothesis for immigrants with an acceptable level of proficiency in the target language should be no or only a very inconsequential amount of bias in safety suitability tests, and no differential prediction, as the language component in the safety suitability tests is minimal.

Can Workforce Diversity be Improved?

Their lower scores on safety suitability tests and cognitive tests lead to less immigrants being hired. The policy concerning safety at Dutch Railways is clear: Lowering standards for safety is simply out of the question, as it would lead to a higher risk of accidents. Moreover, passengers would be less than pleased with riding in trains driven by people selected on the basis of positive discrimination.

How may workforce diversity be improved without lowering standards for safety? A possibility is to redesign the job so that the demands are less. This could be done, for instance, by constructing adjustments of the safety systems specifically aimed at compensating individual's weak points in their profile of safety skills or cognitive skills. Of course, these adjustments need not only apply to immigrants but can be applied generally. Another possibility would be to carefully check the level of safety suitability required. An immigrant applying for a position of train driver may not have the high level of safety suitability required, but his level of safety skills may still be sufficient for meeting middle-level safety requirements.

Still another possibility is to explore whether Ackerman's (1986) model of skill aquisition can be extended to safety skills acquisition. The theory states that a functional equivalence exists between three broad classes of cognitive abilities and three phases of skill aquisition. The controlled phase of skill aquisition is associated with demands on general ability or *g*. The controlled and automatic phase is associated with demands on perceptual-speed abilities. The automatic phase is associated with demands on perceptual-motor abilities. New studies may show that the biggest handicap for the immigrant job applicants in this study is their low mean level of *g*, as it may be linked to success during the first phase of safety skills acquisition. It may be that the second phase of safety skills aquisition is most strongly influenced by perceptual-speed abilities; for Surinamese and Antilleans, this may again imply a handicap, albeit much less stronger than their handicap because of lower mean *g*, for, when they are matched for *g* with majority group applicants, they have somewhat lower perceptual-speed scores. On the other hand, Turks and North Africans have higher perceptual speed, which may give them a relative advantage in this phase. It may be that the final phase of safety skills acquisition is characterized by a dependence upon perceptual-motor abilities; this may again imply a handicap for immigrants, for when they are matched for *g* with majority group applicants, they all show lower psychomotor abilities. So, a possible route to improved workforce diversity is having more immigrants with the required safety skills by extending the standard training. To what extent such an extended training would be successful remains an empirical question. Of course, an extended training may not only be offered to immigrants but also to Dutch, most likely resulting in a diminishment of the effect of increasing workforce diversity.

New Ways of Using the Scores on Safety Tests

All safety tests consist of different parts, and generally the last parts are somewhat less strongly *g* loaded. The question then becomes whether the measurement of abilities should be most strongly based on the scores at the beginning, the middle, or at the end of the tests. Although safety-suitability is conceptually linked to *g*, and one would expect a relation with intelligence, it could be that the cognitive component of the safety-suitability tests is too strong, because test takers are still in the process of trying to understand the test. It may be hypothesized that the later parts of the test, when test takers are more used to the way the test works, provide a better prediction of attention and perceptual-motor skills of experienced incumbents. Ackerman (1990) shows that training on perceptual-motor tests results in higher criterion-related validities after this training, but that training on perceptual-ability tests results in reduced criterion-related validity.

Allowing safety-suitability test scores to reach asymptotic values provides information that may be used as a predictor of the final level of skill on the job, next to *g* and several narrow cognitive abilities. However, it is unrealistic to believe that any selecting agency would bear the cost of repeated administration over several days.

Final Conclusion

The increase of safety jobs and the increasing number of immigrants pose challenges for selection psychologists. Continued use of safety suitability tests is required to keep the number of safety accidents at a minimum. There is very little indication of test bias in safety suitability tests.

Acknowledgements

The Training and Recruitment Foundation of Dutch Railways provided the main part of the financial support for this research. The authors thank two anonymous reviewer for their detailed and constructive criticism of an earlier version of this paper.

References

Ackerman, P.L. (1986) Individual differences in information processing. An investigation of intellectual abilities. *Intelligence*, **10**, 101–139.

Ackerman, P.L. (1987) Individual differences in skill learning: An integration of psychometric and information processing skills. *Psychological Bulletin*, **102**, 3–27.

Ackerman, P.L. (1990) A correlational analysis of skill specifity. Learning, abilities, and individual differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **16**(5), 883–901.

Adams, J.A. (1957) The relationship between certain measures of ability and acquisition of a psychomotor response. *Journal of General Psychology*, **56**, 121–134.

Anderson, J.R. (2000) *Cognitive psychology and its implications* (5th Edn). New York: Freeman.

Arthur, W.J., Barrett, G.V. and Alexander, R.A. (1991) Prediction of vehicular accident involvement. A meta-analysis. *Human Performance*, **42**, 89–105.

Bentler, P.M. (1989) *EQS structural equations program manual*. Los Angeles, CA: BMDP Statistical Software.

Boomsma, L. and Bosch, F. (1978) *De Groepsbourdon [The Group Bourdon]*. Utrecht, the Netherlands: Dutch Railways.

Boomsma-Suerink, J.L. (1985) *PAQ. Analyse van functies in psychologische kenmerken [PAQ. Analysis of jobs in psychological characteristics]*. Utrecht, the Netherlands: Dutch Railways.

Bosch, F. and Schoonman, W. (1982) *Veiligheidstests & betrouwbaarheid. De psychometrische betrouwbaarheid van veiligheidstests by NS [Safety suitability tests & reliability. The psychometric reliability of safety suitability tests at Dutch Railways]*. Utrecht, the Netherlands: Dutch Railways.

Bukasa, B. and Wenninger, U. (1985) Beziehung zwischen verkehrspsychologischen Testverfahren und Kriterien des Fahrverhaltens unter Verwendung einer Fahrertypologie [Relations between scores on psychological traffic tests and criteria of driving behavior while using a driver classification]. *Zeitschrift für Verkehrsicherheit*, **31**, 80–85.

Carroll, J.B. (1993) *Human cognitive abilities. A survey of factor-analytic studies*. Cambridge: Cambridge University Press.

Evers, A., van Vliet-Mulder, J.C. and ter Laak, J. (1992) *Documentatie van tests en testresearch in Nederland [Documentation of tests and test research in the Netherlands]*. Assen, the Netherlands: Van Gorcum.

Fletcher, S. and Geary, B. (1993). Predicting the successful trainee train driver. Paper presented at the British Psychological Society Annual Conference, 2–5 April 1993, Blackpool.

Gorsuch, R.L. (1983) *Factor analysis*. Hillsdale, NJ: Lawrence Erlbaum.

Gottfredson, L.S., Arvey, R.D., Bouchard, T.J., Carroll, J.B., Cohen, D.B., Dawis, R.V., Detterman, D.K., Dunnette, M., Eysenck, H., Feldman, J., Fleishman, E.A., Gilmore, G.C., Gordon, R.A., Greene, R.L., Haier, R.J., Hardin, G., Hogan, R.J., Horn, J.M., Humphreys, L.G., Hunter, J.E., Itzkoff, S.W., Jackson, D.N., Jenkins, J.J., Jensen, A.R., Kaufman, A.S., Kaufman, N.L., Keith, T.Z., Lambert, N., Loehlin, J.C., Lubinski, D., Lykken, D.T., Lynn, R., Meehl, P.E., Osborne, R.T., Perloff, R., Plomin, R., Reynolds, C.R., Rowe, D.C., Rushton, J.P., Sarich, V., Scarr, S., Schmidt, F.L., Schoenfeldt, L.F., Sharf, J.C., Spitz, H., Stanley, J.C., Thiessen, D., Thompson, L.A., Thorndike, R.M., Vernon, P.A. and Willerman, L. (1997) Mainstream science on intelligence. *Intelligence*, **24**, 13–23.

International Union of Railways (1992) Methods and techniques used by psychology departments at European Railways. In *Selection of safety personnel*. Paris: Author .

Jensen, A.R. (1980) *Bias in mental testing*. London: Methuen.

Jensen, A.R. (1985) The nature of the black-white difference on various psychometric tests. Spearman's hypothesis. *The Behavioral and Brain Sciences*, **8**, 193–263.

Jensen, A.R. (1993) Spearman's hypothesis tested with chronometric information-processing tasks. *Intelligence*, **17**, 47–77.

Jensen, A.R. and Weng, L.-J. (1994) What is a good *g*? *Intelligence*, **18**, 231–258.

McCormic, E.J., Jeanneret, P. and Mecham, R.C. (1972) A study of job characteristics and job dimensions as based on the Position Analysis Questionnaires. *Journal of Applied Psychology*, **36**, 347–368. (monograph).

Meyer, D.E., Irwin, D.E., Osman, A.M. and Kounios, J. (1988) The dynamics of cognition and action: Mental processes inferred from speed-accuracy decomposition. *Psychological Review*, **95**, 183–237.

Nunnally, J.C. (1978) *Psychometric theory* (2nd Edn). New York: McGraw-Hill.

Pennock-Román, M. (1992) Interpreting test performance in selective admission for Hispanic students. In K.F. Geisinger (Ed.), *Psychological testing of Hispanics* (pp. 99–135). Washington, DC: APA.

Pew, R.W. (1969) The speed-accuracy operating characteristic. *Acta Psychologica*, **30**, 16–26.

Pouw, E. (1991) Individuele verschillen in reactievermogen: Cognitieve aspecten [Individual differences in reaction ability: cognitive aspects]. In H. van der Flier, P.G.W. Jansen and J.N. Zaal (Eds.), *Selektieresearch in de praktijk*. Lisse, the Netherlands: Swets.

Reynolds, B. and Adams, J.A. (1954) Psychomotor performance as a function of initial level of ability. *American Journal of Psychology*, **67**, 268–277.

Reynolds, C.R. and Kaiser, S.M. (1990) Bias in assessment of suitability. In C.R. Reynolds and R.W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence and achievement* (pp. 611–653). New York: Guilford.

Rushton, J.P. (1998) The "Jensen Effect" and the "Spearman–Jensen hypothesis" of black-white IQ differences. *Intelligence*, **26**(3), 217–225.

Rutten, J.W. and Block, J.R. (1976) *The attention diagnostic method. A test manual*. New York: Instrumental Psychological Methods.

Salgado, J.F., Anderson, N., Moscoso, S., Bertua, C. and de Fruyt, F. (2003) International validity generalization of GMA and cognitive abilities as predictors of work behaviours: A European contribution and comparison with American findings. *Personnel Psychology*, **56**, 573–605.

Schmidt, F.L. and Hunter, J.E. (1998) The validity and utility of selection methods in personnel psychology. Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, **124**(2), 262–274.

Schmidt, F.L., Ones, D.S. and Hunter, J.E. (1992) Personnel selection. *Annual Review of Psychology*, **43**, 627–670.

Schoonman, W. and Bosch, F. (1981) *Test & oordeel. Onderzoek naar de selektie van autobuschauffeurs. [Test & assessment. A study into the selection of bus drivers]*. Utrecht, the Netherlands: Bedrijfspsychologie, Dutch Railways.

Spearman, C. (1927) *The abilities of man: Their nature and measurement*. London: MacMillan.

te Nijenhuis, J., Evers, A. and Mur, J.P. (2000) The validity of the Differential Aptitude Test for the assessment of immigrant children. *Educational Psychology*, **20**, 99–115.

te Nijenhuis, J. and van der Flier, H. (1997) Comparability of GATB scores for immigrants and majority group members: Some Dutch findings. *Journal of Applied Psychology*, **82**, 675–687.

te Nijenhuis, J. and van der Flier, H. (1999) Bias research in the Netherlands. Review and implications. *European Journal of Psychological Assessment*, **15**(2), 165–175.

te Nijenhuis, J. and van der Flier, H. (2000) Differential prediction of immigrant versus majority group training performance using cognitive ability and personality measures? *International Journal of Selection and Assessment*, **8**, 54–60.

te Nijenhuis, J. and van der Flier, H. (2001) Group differences in mean intelligence for the Dutch and third World immigrants. *Journal of Biosocial Science*, **33**, 469–475.

te Nijenhuis, J. and van der Flier, H. (2002) The correlation of *g* with attentional and perceptual-motor ability tests. *Personality and Individual Differences*, **33**, 287–297.

te Nijenhuis, J. and van der Flier, H. (2003) Immigrant-Majority group differences in cognitive performance: Jensen effects, cultural effects, or both? *Intelligence*, **31**, 443–459.

te Nijenhuis, J., van der Flier, H. and van Leeuwen, L. (1997) Comparability of personality test scores for immigrants and majority group members: Some Dutch findings. *Personality and Individual Differences*, **23**, 849–859.

te Nijenhuis, J., van der Flier, H. and van Leeuwen, L. (2003) The use of a test for neuroticism, extraversion, and rigidity for Dutch immigrant job-applicants. *Applied Psychology: An International Review*, **52**, 630–647.

van der Flier, H. and Boomsma-Suerink, J.L. (1994) *Handboek GATB. [GATB manual.]*. Amsterdam: Stichting GATB – Research.

van der Flier, H. and Schoonman, W. (1988) Railway signals passed at danger. Situational and personal factors underlying stop signal abuse. *Applied Ergonomics*, **19**, 135–141.

van der Flier, H., Schoonman, W. and Pouw, E. (1993) Contribution of psychological tests to railway safety. In *Selection of safety personnel. The 1st European Congress of Railway Psychology*. Paris: International Union of Railways.

van Drie, H. and Schoonman, W. (1993) *NS CAT Handboek [Manual NS CAT]*. Utrecht, the Netherlands: Kwaliteitsbureau, Dutch Railways.

Vos, P.G. (1988) *Bourdon Vos test. Handleiding [Manual Bourdon-Vos test]*. Lisse, the Netherlands: Swets.

